Information Retrieval Tutorial 1 10/08/2018

1. Write the complexity of retrieving query
a)“Brutus AND (NOT Caesar)”
b)”Brutus OR NOT Caesar”
by boolean retrieval merging? Assume size of posting list of brutus and caesar be x and y respectively.

Ans. a) O(x+y) b) O(N)

2. Can skip pointers can be used in every case? Why not?

Ans. Nope, (x or y)

3. Calculate the number of comparisons for merging the following:

4 6 10 12 14 16 18 20 22 32 47 81 120 122 157 180
47

a. Normal postings lists
b. Skip pointers

Ans. a) 11 b) 6

* **4**. Derive the complexity for positional index?

L=Total number of occurrences of two terms in document.
K clause phrase
m and n = size of postings list of both words.

Ans. O((m+n)L)

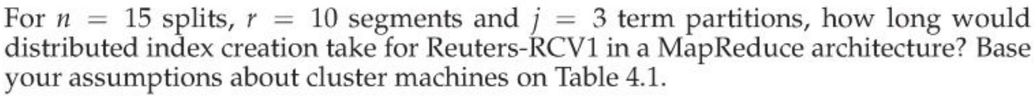
5. State the problem of using conjunction of bigrams with a example.

Ans. mon*h will falsely match moonish.

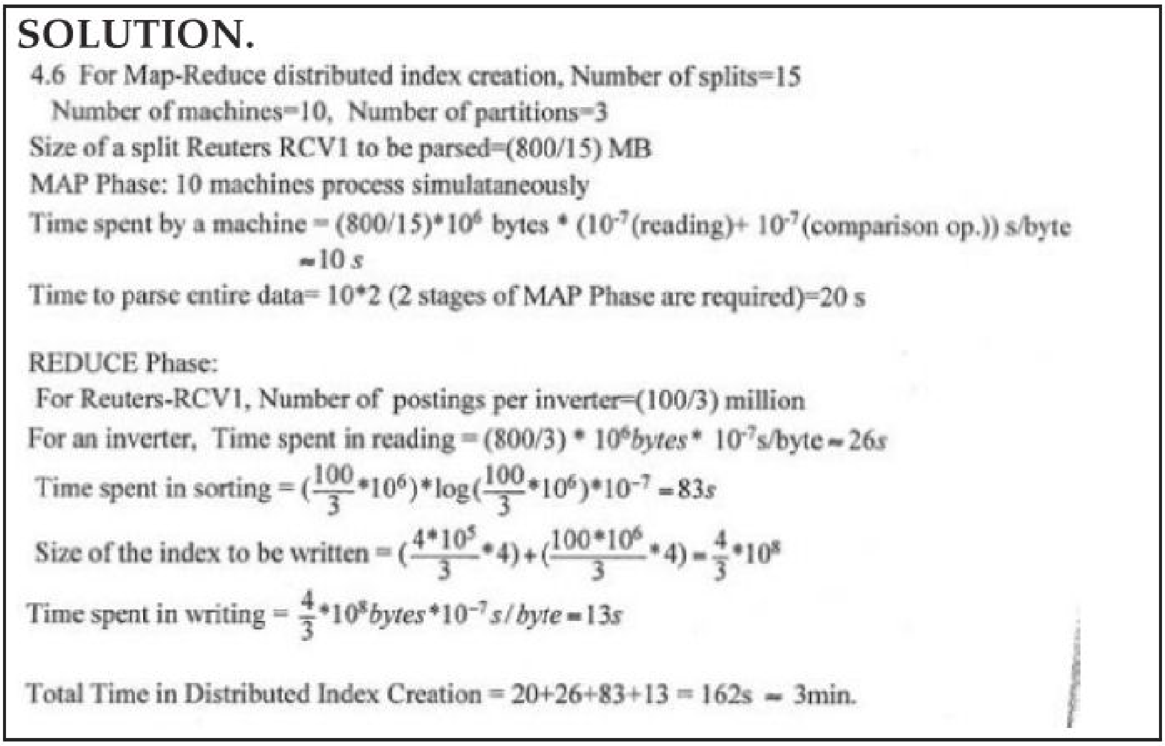
6. Jaccard Coefficient between bord and sordid (bigram)

Ans. 2/6

7.



For $n = 15$ splits, $r = 10$ segments and $j = 3$ term partitions, how long would distributed index creation take for Reuters-RCV1 in a MapReduce architecture? Base your assumptions about cluster machines on Table 4.1.



SOLUTION.

4.6 For Map-Reduce distributed index creation, Number of splits=15
Number of machines=10, Number of partitions=3

Size of a split Reuters RCV1 to be parsed=(800/15) MB

MAP Phase: 10 machines process simulataneously

Time spent by a machine = $(800/15)*10^6$ bytes * (10^{-7}(reading)+ 10^{-7}(comparison op.)) s/byte
$\approx 10\ s$

Time to parse entire data= 10*2 (2 stages of MAP Phase are required)=20 s

REDUCE Phase:

For Reuters-RCV1, Number of postings per inverter=(100/3) million

For an inverter, Time spent in reading = $(800/3) * 10^6 bytes * 10^{-7}$s/byte $\approx 26s$

Time spent in sorting $= (\frac{100}{3}*10^6)*\log(\frac{100}{3}*10^6)*10^{-7} = 83s$

Size of the index to be written $= (\frac{4*10^5}{3}*4)+(\frac{100*10^6}{3}*4) \approx \frac{4}{3}*10^8$

Time spent in writing $= \frac{4}{3}*10^8 bytes*10^{-7} s/byte \approx 13s$

Total Time in Distributed Index Creation = 20+26+83+13 = 162s $\approx$ 3min.